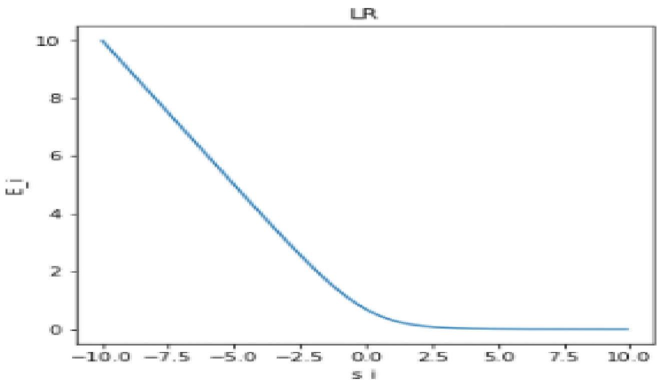
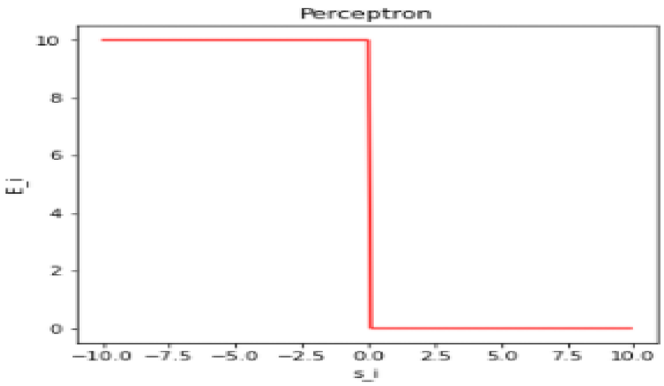
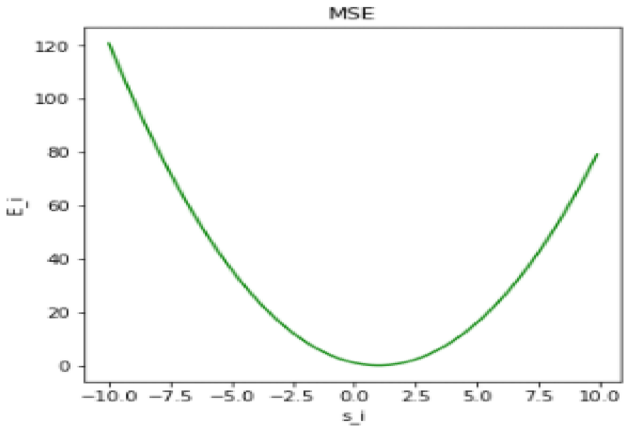
1) (a)


LR
10
8
6
4
2
0
E_i
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
s_i

(b)


Perceptron
10
8
6
4
2
0
E_i
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
s_i

(c)


MSE
120
100
80
60
40
20
0
E_i
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
s_i

a)(i)
Dataset1:

	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	9.365356666783853e-28	480.897802195002
	w	[-7.01477582 3.20265861 -2.01056618 4.61891474 -8.48679639 5.34513234 -1.36854253 -20.00142649 13.2641012 3.11232438]			
		$l_1(w)$ =68.42523867301658	$l_\$(w)$ =27.80269624088004	Spars=0	
LASSO	2	120.35818995011604	114.19046963120057	14.107883849225086	233.3835984423175
	w	[0.12578696 2.26001059 0. -3.34237423 -0. 5.01163416 0. -5.93509725 -0. 1.43300028]			
		$l1(w)$ =18.107903459292228	$l\$(w)$ =8.870754296600296	Spars=4	
Ridge	4.5	84.5753041414547	43.97256195045436	21.496921716819042	270.97166281655035
	w	[-0.13458667 2.45289798 -0.20945981 -1.73524002 -1.56346484 2.70527836 2.29260148 -2.95259376 -2.74902616 1.51411343]			
		$l1(w)$ =18.30926250913052	$l\$(w)$ =6.532704072988467	Spars=0	

Dataset2:

	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	86.33661129877157	112.65154328000656
	$\boldsymbol{w}$	[0.43392102 2.397075 0.5682055 -3.87069203 0.8554485 2.25097789 2.04197312 -6.17726984 -1.80441184 1.25424529]			
		$l_1(w)$ =21.654220029500763	$l_{\$}(w)$ =8.613675992794823	Spars=0	
LASSO	1	108.49900700912701	46.86211987945499	88.6600338486532	110.96937070707098
	$\boldsymbol{w}$	[0.49085208 2.30676475 0.29571815 -2.88426941 -0. 2.35888612 1.89720305 -6.34889531 -1.58317001 1.08490682]			
		$l1(w)$ =19.250665701460257	$l\$(w)$ =8.192936198374936	$\text{Spars}=1$	
Ridge	6	107.71919354132224	43.73350356057717	89.14761102319817	111.42028497489187
	$\boldsymbol{w}$	[0.45904623 2.25533004 0.55844399 -2.57037539 -0.32269209 2.23048119 2.05571123 -4.14875114 -3.77234633 1.17294188]			
		$l1(w)$ =19.54611950615056	$l\$(w)$ =7.371538310088321	$\text{Spars}=0$	

Dataset3:

	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	98.21301479827	109.12481315987688
	$\boldsymbol{w}$	[1.71594731 1.90468457 0.41212604 -3.17204863 0.25311452 4.87289258 -0.25297342 -8.71299177 0.80571383 0.89176542]			
		$l_1(w)$ =22.994258103832298	$l_{\$}(w)$ =10.864521591717066	Spars=0	
LASSO	-1.5	100.13262659725935	11.712019479000341	98.47295449776546	109.27565304657898
	$\boldsymbol{w}$	[1.69887908 1.88246962 0.36694448 -2.90839199 -0. 4.60540281 0. -7.90046258 -0. 0.86234901]			
		$l1(w)$ =20.22489957013189	$l\$(w)$ =9.969652150067558	Spars=3	
Ridge	3	101.08532318473706	12.135068922700214	98.24920712781592	108.87756214285892
	$\boldsymbol{w}$	[1.71517391 1.90359855 0.41126392 -3.15279029 0.2340847 4.79880832 -0.18034491 -8.23585846 0.32967973 0.89031463]			
		$l1(w)$ =21.851917433305715	$l\$(w)$ =10.417357069682891	Spars=0	

ii)(1) as expected and observed in Dataset1, the test error has significantly improved after regularization using Ridge and Lasso. With no regularization, the model overfitted the data and had almost zero training error, but it generalized poorly to the test data. After regularization, the generalization performance has improved and Lasso performed slightly better than Ridge, probably because of the sparsity.

The test error also reduced with increasing the training points, i.e., increasing the data points might be thought of as having an effect of regularization. For Dataset 2 and 3, the regularization didn't have much effect on generalization performance because the model, before regularization, didn't overfit the training data, i.e., it already had a decent generalization

(2) in Dataset1, the norm of w has significantly reduced after regularization, because the training points were few and the model was highly wiggly and overfit to these few data, and hence high values of the coefficients. After regularization, we constrain the value of the coefficients, and hence we get a lower norm of w.

In Dataset 2 and 3, the norm of w didn't reduce significantly after regularization because the model wasn't highly over fit to the training data, and it already had a good generalization performance. As observed, increasing the number of training points has an effect of regularization, it reduces the variance of the model. So, the model's coefficients, before regularization, were not large. However, there was a slight reduction in the norm after regularization. In Daraset3, the norm in Lasso was lower than that of Ridge because of the feature selection property in Lasso. while in Dataset2, Ridge's norm was slightly lower, there were no feature selection in Lasso.

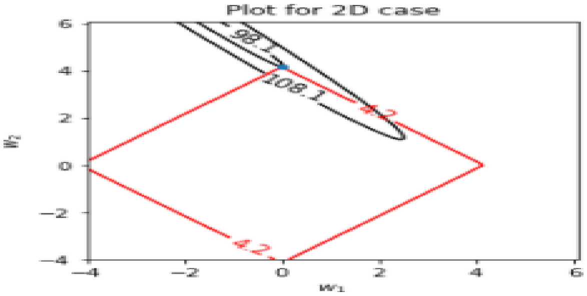
(3) as observed in Datasets 1 and 3, some of the coefficients completely diminished, and that is because of the sharp edges of the constraint region of the l1 norm function, so there is a higher probability that the constrained optimization will be satisfied in a point where some coefficients are zero, where it is not the case in Ridge l2 norm function.

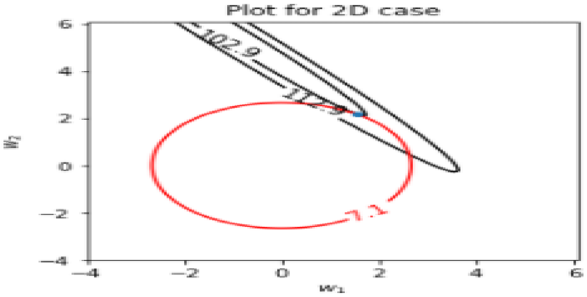
As we increase lambda, i.e., reducing the constraint region, we are forcing the coefficients to be around zero (more sparsity). As observed in Dataset1, Lasso's lambda was higher than that of the other two datasets, and hence we got the highest sparsity.

A similar argument, increasing the number of training points have the effect of regularization, and hence the coefficients get smaller as we increase training points as observed in the tables above. When we increased the training points from 100 to 1000, the coefficients reduced and we had more sparsity.

b) (i)
Dataset4

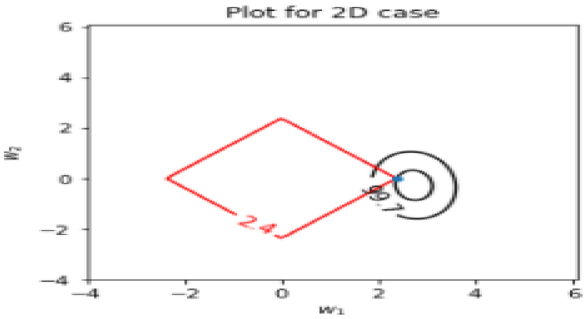
	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	95.38019904643616	163.48761227397384
	w	[6.77265711 -2.4928513 7.23801612]			
		$l_1(w)$ =16.503524531665285	$l_\$(w)$ =10.221157922129624	Spars=0	
LASSO	0.5	161.54595243084395	118.40018863689559	98.09379261871835	139.04677640667703
	w	[5.40758378 0. 4.15574202]			
		$l1(w)$ =9.563325800929503	$l\$(w)$ =6.8199819709990335	Spars=1	
Ridge	4	155.9026878419882	121.95637904426236	102.92593770240384	127.90734292133521
	w	[4.61680832 1.55139276 2.16748412]			
		$l1(w)$ =8.335685196859789	$l\$(w)$ =5.331015470702115	Spars=0	

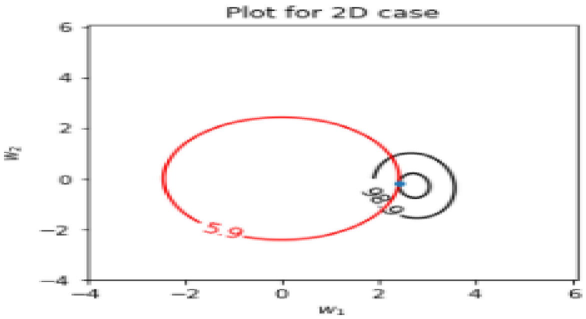




Dataset5

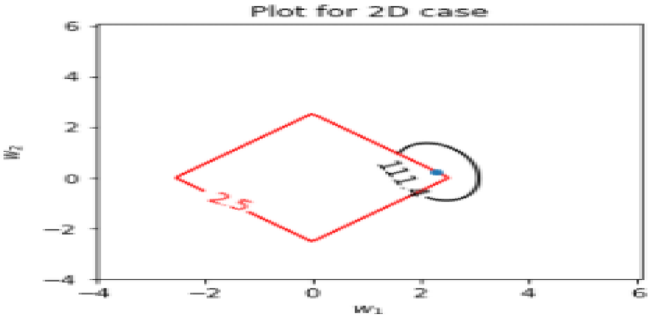
	Model selection			Performance	
	Best param $\log_\$ \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	87.12261767437649	114.70433167932686
	$\boldsymbol{w}$	[4.05510307 2.74884213 -0.29784002]			
		$l_1(w)$ =7.101785217832999	$l_\$(w)$ =4.908024311927914	Spars=0	
LASSO	2.5	99.29431292631122	51.68578141571634	89.72181222213327	103.6350573042968
	$\boldsymbol{w}$	[3.6870101 2.37618863 -0.]			
		$l1(w)$ =6.06319872716057	$l\$(w)$ =4.386378444558678	Spars=1	
Ridge	6	104.57663846874671	52.505687348248756	88.86631773419619	106.00172830717882
	$\boldsymbol{w}$	[3.79650626 2.41954902 -0.18839922]			
		$l1(w)$ =6.404454500561942	$l\$(w)$ =4.505904073798311	Spars=0	

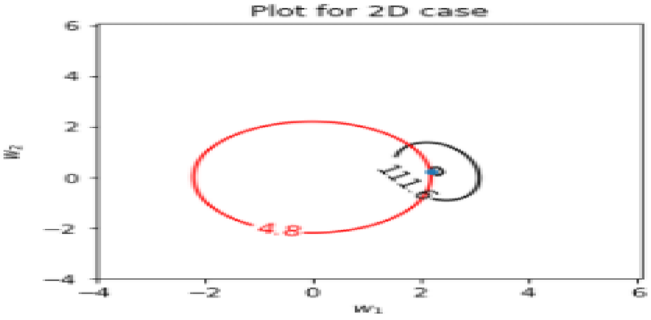




Dataset6

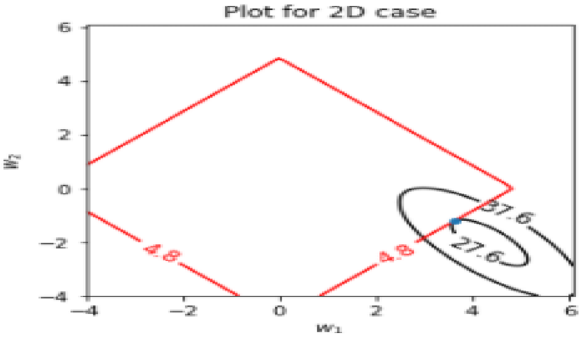
	Model selection			Performance	
	Best param $\log_\$ \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	101.35833777888637	101.4457093370796
	$\boldsymbol{w}$	[1.21077089 2.30240071 0.23152547]			
		$l_1(w)$ =3.7446970691535286	$l_\$(w)$ =2.611631526967983	Spars=0	
LASSO	-10	110.87841697917038	27.339722364878934	101.35833791468285	101.44595275106839
	$\boldsymbol{w}$	[1.2107991 2.30236633 0.23141756]			
		$l1(w)$ =3.7445829866482643	$l\$(w)$ =2.6116047303958667	Spars=0	
Ridge	6.5	110.4223292969328	29.169609209886165	101.58772152431895	101.3082994749359
	$\boldsymbol{w}$	[1.22395431 2.18504022 0.24383607]			
		$l1(w)$ =3.6528306061457267	$l\$(w)$ =2.516330851694283	Spars=0	

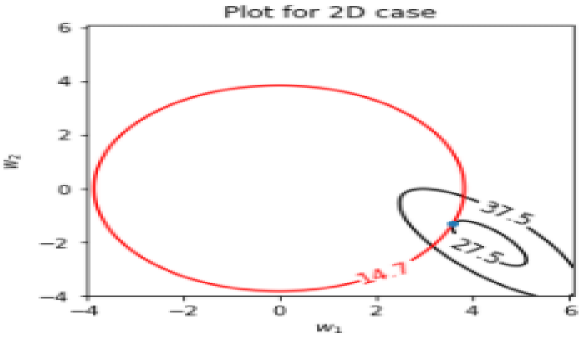




Dataset7

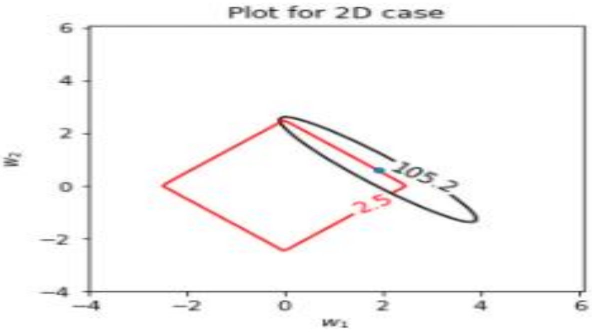
	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	25.417551693358597	116.51141337592613
	$\boldsymbol{w}$	[1.6193184 4.35846137 -2.05316003]			
		$l_1(w)$ =8.030939797861013	$l_{\$}(w)$ =5.082700433707279	Spars=0	
LASSO	0.5	61.91109702592708	51.81377203033446	27.626827689407435	105.78138791318894
	$\boldsymbol{w}$	[1.44203979 3.63303055 -1.21671938]			
		$l1(w)$ =6.291789728863818	$l\$(w)$ =4.093750824293747	Spars=0	
Ridge	2.5	54.831782966414174	53.3268705194972	27.50581291945118	107.48904375692914
	$\boldsymbol{w}$	[1.59787074 3.60182849 -1.32250097]			
		$l1(w)$ =6.522200204625349	$l\$(w)$ =4.1563647811315585	Spars=0	

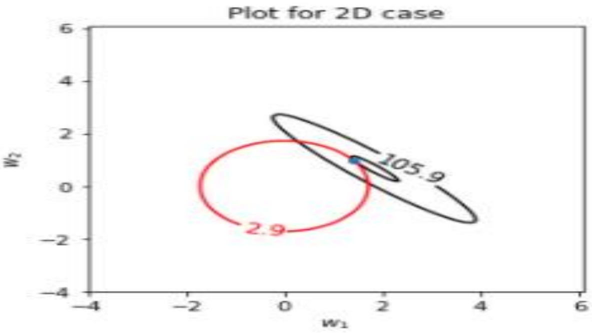




Dataset8

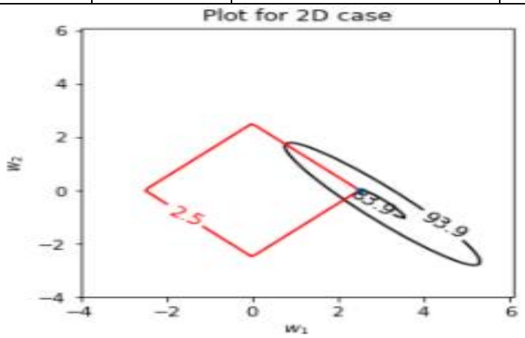
	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	95.15432277075584	109.24257017878496
	$\boldsymbol{w}$	[3.58068323 1.91863829 0.60434473]			
		$l_1(w)$ =6.103666255666341	$l_{\$}(w)$ =4.107030295969648	Spars=0	
LASSO	-0.5	112.25807065948239	37.431892972962494	95.17784459618632	109.38954465644765
	$\boldsymbol{w}$	[3.53520373 1.89335513 0.59624288]			
		$l1(w)$ =6.0248017392607816	$l\$(w)$ =4.054375985065424	Spars=0	
Ridge	6	108.75084091972133	41.107302706474876	95.90349733909385	110.51198296942115
	$\boldsymbol{w}$	[3.20005109 1.41174703 0.97725369]			
		$l1(w)$ =5.589051809257595	$l\$(w)$ =3.631581118965082	Spars=0	

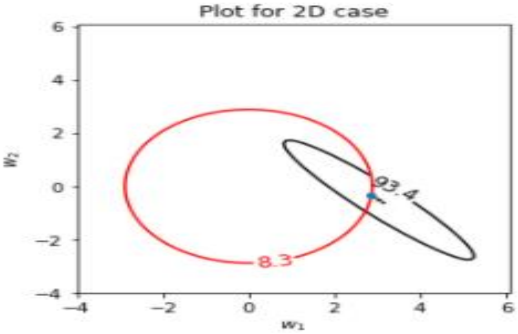




Dataset9

	Model selection			Performance	
	Best param $\log_{\$} \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	83.32401525715771	111.41165530589785
	$\boldsymbol{w}$	[4.11404128 3.04009919 -0.51630424]			
		$l_1(w)$ =7.670444709675474	$l_\$(w)$ =5.141411166840955	Spars=0	
LASSO	0	88.79325849350754	20.28235442339538	83.90836860287891	109.50398425369923
	$\boldsymbol{w}$	[4.10151715 2.50447238 -0.]			
		$l1(w)$ =6.60598952750575	$l\$(w)$ =4.80570752521165	Spars=1	
Ridge	3.5	89.34875318191997	19.68690482275225	83.38817183255067	110.60800347490775
	$\boldsymbol{w}$	[4.1101324 2.86266206 -0.34484516]			
		$l1(w)$ =7.3176396252729825	$l\$(w)$ =5.020651415930633	Spars=0	





iii) 1- In Lasso, when the minimum error that satisfies the constraint is on the edge of the constraint region, we will have sparsity, i.e., one of the feature coefficients will be 0. Unlike Ridge, where it is highly unlikely that there will be sparsity because the nature and smoothness of the l2 norm constraint region, where there are no edges.

2- we had a better test error performance after regularization. In rich Datasets and in this case (e.g. Dataset 6), however, the effect of regularization on test performance is not considerable, because the model wasn't overfit to the data and already had a good generalization performance. We can infer from the plots that the Train MSE for the unregularized case is increasing until it satisfies the constraint region, and hence at this point, this is the new regularized MSE.

3- There were no feature selections performed by Lasso in these datasets. However, when we heavily increase the number of training points (Dataset 9), the variance of the model reduced and one of the coefficients diminished.